Biometrics 74, 1104–1111
September 2018

DOI: 10.1111/biom.12834

New Semiparametric Method for Predicting High-Cost Patients

Adam Maidman * **and Lan Wang** **

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
* *email:* maidm004@umn.edu
** *email:* wangx346@umn.edu

Summary. Motivated by the Medical Expenditure Panel Survey containing data from individuals' medical providers and employers across the United States, we propose a new semiparametric procedure for predicting whether a patient will incur high medical expenditure. Problems of the same nature arise in many other important applications where one would like to predict if a future response occurs at the upper (or lower) tail of the response distribution. The common practice is to artificially dichotomize the response variable and then apply an existing classification method such as binomial regression or a classification tree. We propose a new semiparametric prediction rule to classify whether a future response occurs at the upper tail of the response distribution. The new method can be considered a semiparametric estimator of the Bayes rule for classification and enjoys some nice features. It does not require an artificially dichotomized response and better uses the information contained in the data. It does not require any parametric distributional assumptions and tends to be more robust. It incorporates nonlinear covariate effects and can be adapted to construct a prediction interval and hence provides more information about the future response. We provide an `R` package `plaqr` to implement the proposed procedure and demonstrate its performance in Monte Carlo simulations. We illustrate the application of the new method on a subset of the Medical Expenditure Panel Survey data.

Key words: Conditional quantile; Expenditure prediction; High-cost patient; Partially linear additive model; Semiparametric regression; Upper tail.

1. Introduction

Our work is motivated by the Medical Expenditure Panel Survey (MEPS). MEPS contains nationally representative data from individuals' medical providers and employers across the United States and is the most complete source of data on cost and use of healthcare. Medical expenditure is the total cost of all healthcare services paid out-of-pocket, by insurance, or by other sources, excluding the purchase of over-the-counter drugs. It is well known that a large proportion of health care costs is concentrated in a relatively small portion of patients. The estimated total medical expenditure in the United States in 2008 was about US $1.15 trillion; yet just 10% of patients were responsible for about 64% of the total spending (Schoenman and Chockley, 2011). Identifying high-cost patients prospectively is an important step toward controlling future health care costs. Ahn et al. (2015) found that enrollment in a community program designed to reduce hospital visits can save individuals about US $ 700–US $1000. The per-person cost of enrollment varies with population size, but can be between US $100 and US $350. Thus, the cost of failing to enroll an individual who will benefit from the program can range from US $350 to US $900, while the cost of enrolling someone who will not benefit can range from US $100 to US $350.

In this article, we propose a new semiparametric prediction procedure using training data from the past one or two years to classify a patient's next-year expenditure into the class of "high-cost" or "not-high-cost." A threshold value c determined by a field expert, typically corresponding to a high quantile of the expenditure distribution, separates the two classes. This problem differs from the traditional classification problem in two important aspects. First, the actual values of the response variable on a continuous scale are available in the training data set, not solely class labels. Second, the two classes are severely imbalanced with high-cost patients in the minority. Ignoring the first issue results in efficiency loss; while ignoring the second issue results in a classification rule with low sensitivity, that is, low probability of identifying the high-cost patients. An additional difficulty inherent in expenditure data is skewness and heteroskedasticity which pose challenges for statistical analysis (Zhou et al., 2001) and prediction at the tails of the distribution.

A popular approach in the literature for predicting if a new subject will be located in the tails of the response distribution relies on binomial regression using a logistic link function, for example, Fleishman and Cohen (2010), Meenan et al. (1999), and Hosmer Jr and Lemeshow (2004). Other link functions such as the complementary log–log function may be used as well. Given the threshold c, the binomial regression approach first artificially discretizes medical expenditure by assigning a value of 1 if the expenditure is greater than c and 0 otherwise. A binomial regression model is then fit to the 0–1 response data and a new patient can be classified as high-cost if his or her predicted probability of being a high-cost patient is more likely than not. By artificially dichotomizing the response, binomial regression results in efficiency loss and it is not clear whether the artificially modified data satisfy modeling assumptions.

Modeling English inpatient healthcare expenditure using the generalized beta distribution of the second kind and the generalized gamma distribution was found to have potential in predicting tail probabilities (Jones et al., 2015). These methods can suffer from high variability without very large sample sizes. Bertsimas et al. (2008) took algorithmic approaches to predicting future healthcare expenditure using classification trees (Breiman et al., 1984) and clustering algorithms (Kannan et al., 2004). While clustering algorithms are useful for identifying similar groups of patients, they cannot predict if a future patient belongs to a class defined a priori.

If 10% of all patients are high-cost, the naive classification rule that classifies every patient as not-high-cost has merely a 10% error rate. However, it completely misses the minority class of high-cost patients rendering it unsuitable for many applications (Vickers and Elkin, 2006). Let r be the ratio of costs of a false positive (a not-high-cost patient predicted to be high-cost) and a false negative (a high-cost patient predicted to be not-high-cost). Simply taking $r = 1$ can result in classification rules with low sensitivity.

We propose a novel procedure that takes into account the missclassification error costs and leads to increased performance of sensitivity and overall classification. Our procedure uses training data to obtain a semiparametric estimation of the $\left(\frac{1}{1+r}\right)$th conditional quantile function. The classification rule amounts to comparing the $\left(\frac{1}{1+r}\right)$th conditional quantile of the response with the given threshold c. We show that this semiparametric procedure consistently estimates the Bayes rule. The new prediction procedure does not require dichotomization of the response and fully uses the information contained in the expenditure data. It does not require parametric distributional assumptions and is possibly more robust. The proposed procedure can be modified to create prediction intervals for future expenditure yielding richer information. In contrast, binomial regression provides little extra information beyond predicting whether the future expenditure is below or above the threshold.

In Section 2 of the article, we provide background information on the semiparametric quantile regression model, which is a necessary building block of the proposed classification method. Section 3 introduces the new semiparametric classification procedure and its connection to binomial regression. We demonstrate the performance of our new estimator with Monte Carlo simulations in Section 4. Section 5 reports a detailed analysis of MEPS. We conclude with a discussion in Section 6. Numerical results in Section 4 demonstrate that the new classification procedure is better able to correctly classify new patients, particularly high-cost patients, compared to existing parametric and algorithmic procedures.

2. Background

Suppose, we have training data $(Y_i, X_i')'$, $i = 1, \ldots, n$, where the continuous variable Y_i denotes the ith patient's next-year expenditure and $X_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of predictors. Given a new patient (Y^*, X^*) where only X^* is observable, we would like to classify the patient as high-cost or not-high-cost based on the value X^*. In the following, we give a brief introduction to quantile regression which is essential for our new semiparametric procedure for classifying future patients.

The conditional distribution function of Y given X is $F_{Y|X}(y) = P(Y \le y|X)$. For a given $0 < \tau < 1$, the τth conditional quantile of Y given X is defined as $Q_{Y|X}(\tau) = \inf\{t : F_{Y|X}(t) \ge \tau\}$. Q is the generalized inverse of F, such that $F(Q(\tau)) \ge \tau$ and $Q(F(y)) \le y$, with equality holding for absolutely continuous random variables. The conditional median is $Q_{Y|X}(0.5)$. Interpretation of the conditional quantile is straightforward. For example, given the vector of predictors $X = x$ and $\tau = 0.9$, 90% of observations of Y with associated $X = x$ fall below $Q_{Y|X}(0.9)$. A useful property of the quantile function is the invariance property. For any monotone function h, $Q_{h(Y)|X}(\tau) = h\,(Q_{Y|X}(\tau))$; the analog for the conditional mean is not always true, that is, in general $E[h(Y)|X] \ne h(E[Y|X])$. We refer to Koenker (2005) for a comprehensive introduction to quantile regression.

2.1. *Partially Linear Additive Quantile Regression*

Exploratory analysis of MEPS reveals that some covariates have nonlinear effects on the response. More specifically, we write $X_i = (V_i', Z_i')'$, where V_i is a p-vector of covariates with linear effects and $Z_i = (Z_{i1}, \ldots, Z_{iq})'$ is a q-vector of covariates with nonlinear effects. To incorporate nonlinearity, we use the flexible partially linear additive quantile regression model. Given the random sample $(Y_i, X_i')'$, $i = 1, \ldots, n$, the partially linear additive quantile regression model assumes that

$$Q_{Y_i|X_i}(\tau) = V_i'\beta(\tau) + \sum_{k=1}^{q} g_k(Z_{ik}), \tag{2.1}$$

where g_k is an unknown smooth nonparametric function, $k = 1, \ldots, q$. Alternatively, we can write

$$Y_i = V_i'\beta(\tau) + \sum_{k=1}^{q} g_k(Z_{ik}) + \varepsilon_i,$$

where the errors $\{\varepsilon_i\}_{i=1}^n$ are independent and satisfy the quantile constraint $P(\varepsilon_i \le 0|X_i) = \tau$. Due to no assumed parametric distribution for ε_i nor constant variance, quantile regression is an attractive model for modeling skewed and heteroscedastic expenditure data. For identifiability, we assume that $E(g_k(Z_{ik})) = 0$. Semiparametric quantile regression models considered by He and Shi (1996), He et al. (2002), Wang et al. (2009), among others, are useful for incorporating nonlinearity while avoiding the curse of dimensionality.

We approximate the unknown, nonlinear functions g_k with a linear combination of B-spline basis functions. We refer to Schumaker (1981) for the construction of B-spline basis functions. Without loss of generality, we assume the covariates Z_{ik} are standardized to be in the interval [0,1]. Let $\pi(t) = (b_1(t), \ldots, b_{k_n+l+1}(t))'$ denote a vector of normalized B-spline basis functions of order $l+1$ with k_n quasi-uniform internal knots on $[0, 1]$. Then $g_k(Z_{ik})$ can be approximated by $\pi(Z_{ik})'\xi_k$, where ξ_k are estimated from the data, $k = 1, \ldots, q$. The B-spline approximation is flexible and computationally efficient. For simplicity, we use the same number of basis functions for all nonparametric components, but this is not necessary in practice.

The estimator for the partially linear additive quantile regression model is

$$\left\{\hat{\beta}(\tau), \widehat{\xi}_1(\tau), \ldots, \widehat{\xi}_q(\tau)\right\} = \underset{\{\beta,\xi_1,\ldots,\xi_q\}\in\mathbb{R}^{p+(k_n+l+1)q}}{\arg\min} \sum_{i=1}^{n} \rho_\tau \left[Y_i - \left\{ V_i'\beta + \sum_{k=1}^{q} \pi(Z_{ik})'\xi_k \right\} \right],$$

with loss function $\rho_\tau(u) = u(\tau - I\{u < 0\})$. The optimization problem can be effectively solved by linear programming (Koenker and d'Orey, 1987). The estimator for the nonparametric function g_k is

$$\hat{g}_k(Z_{ik}) = \pi(Z_{ik})'\hat{\xi}_k(\tau) - n^{-1}\sum_{i=1}^{n} \pi(Z_{ik})'\hat{\xi}_k(\tau), \tag{2.2}$$

for $k = 1, \ldots, q$; where the centering is the sample analog of the identifiability condition $E[g_k(Z_{ik})] = 0$. In the sequel, we will omit the dependence on τ in notation when the quantile level of interest is clear from context. The asymptotic theory of the estimator is investigated in Sherwood and Wang (2016). For consistency, it is required that $k_n \to \infty$, but in practice usually the choice of a small integer works well.

3. New Semiparametric Prediction Procedure

3.1. *Bayes Rule for Classification*

A patient is considered as high-cost if his or her next-year medical expenditure, denoted by Y, is greater than a predetermined threshold c. We consider a loss function that allows for unequal weighting of a false positive and a false negative. For a new patient with covariates X^*, let $\phi(X^*) \in \{1, -1\}$ be the prediction: -1 for not-high-cost and 1 for high-cost. The loss function of the decision rule $\phi(X^*)$ is

$$L(\phi(X^*)) = \begin{cases} r^{-1}, & \text{if } \phi(X^*) = -1 \text{ and } Y^* > c, \\ 1, & \text{if } \phi(X^*) = 1 \text{ and } Y^* \le c, \\ 0, & \text{otherwise.} \end{cases}$$

Without loss of generality, the cost of a false positive is 1. Hence, r is the ratio of the cost of a false positive to a false negative. Taking $r = 1$ can result in classification rules with low sensitivity. Smaller ratios that weight the cost of a false negative heavier than that of a false positive (for example, ratio of 4:1) result in classification rules with higher sensitivity. The ratio can be supplied by field experts or estimated from a pilot study. Similar to the threshold c, the ratio r is driven by the domain of application, not by the data.

The Bayes rule for classification minimizes the expected weighted 0–1 loss function,

$$E\left[L(\phi(X^*))\right] = I\left(\phi(X^*) = 1\right)\left[1 - P(Y^* > c \mid X^*)(1 + r^{-1})\right] + r^{-1}P(Y^* > c \mid X^*).$$

It is straightforward to show that the decision rule $\phi(X^*)$ that minimizes $E\left[L(\phi(X^*))\right]$ is given by

$$\phi(X^*) = \begin{cases} 1, & \text{if } P(Y^* > c \mid X^*) > \frac{r}{1+r}, \\ -1, & \text{if } P(Y^* > c \mid X^*) \le \frac{r}{1+r}. \end{cases} \tag{3.1}$$

3.2. *The New Prediction Method*

We want to classify a new patient with known predictors X^* as high-cost if $Y^* > c$. Note that the Bayes rule classifies a new patient with covariates $X = x^*$ as high-cost if $P\left(Y^* > c \mid x^*\right) > \frac{r}{1+r}$. When $r = 1$ (equally weighted errors), the patient is classified as high-cost if $P\left(Y^* > c \mid x^*\right) > P\left(Y^* \le c \mid x^*\right)$.

Our new approach can be viewed as a semiparametric method for estimating the Bayes rule without directly estimating the class probability $P\left(Y^* > c \mid x^*\right)$. This is based on the important observation that

$$\operatorname{sign}\left[P(Y^* > c \mid X^*) - \frac{r}{1+r}\right] = \operatorname{sign}\left[Q_{Y^*|X^*}\left(\frac{1}{1+r}\right) - c\right]. \tag{3.2}$$

This equivariance suggests that we can estimate the Bayes rule by obtaining a semiparametric estimate of $Q_{Y^*|X^*}\left(\frac{1}{1+r}\right)$ and comparing our estimate to the given threshold c. The approach is semiparametric in the sense that it does not assume a specific parametric distribution model for Y given X.

The classification rule is constructed from the training data $(Y_i, X_i')'$, $i = 1, \ldots, n$, and the observed vector of predictors $X^* = (V^{*'}, Z^{*'})'$ for the new patient in the following three step algorithm.

(1) Fit model (2.1) on the training data and obtain $\hat{\beta}$ and $\hat{g}_k$, $k = 1, \ldots, q$ for $\tau = \frac{1}{1+r}$.
(2) For the new patient, we estimate $\widehat{Q}_{Y^*|X^*}\left(\frac{1}{1+r}\right) = V^{*'}\hat{\beta} + \sum_{k=1}^{q} \hat{g}_k(Z_k^*)$.
(3) Make the prediction: If $\widehat{Q}_{Y^*|X^*}\left(\frac{1}{1+r}\right) > c$, we classify the new patient as high-cost; otherwise, we classify him or her as not-high-cost.

Note that (3.2) is stated for the unknown population conditional quantile function $Q_{Y^*|X^*}$. In Web Appendix A, we show that the proposed semiparametric procedure consistently estimates the Bayes rule,

$$\operatorname{sign}\left[\widehat{Q}_{Y^*|X^*}(\tau) - c\right] = \operatorname{sign}\left[P(Y^* > c \mid X^*) - \frac{r}{1+r}\right] + o_p(1). \tag{3.3}$$

The crux of the derivation is to show that

$$\sup_{z\in[0,1]} \left| \sum_{k=1}^{q} \widehat{g}_k(z) - \sum_{k=1}^{q} g_k(z) \right| = o_p(1). \tag{3.4}$$

In other words, the difference between the estimated and true values of the nonlinear functions goes to zero uniformly as the sample size increases to infinity. This property ensures $\widehat{Q}_{Y^*|X^*}(\tau)$ accurately estimates $Q_{Y^*|X^*}(\tau)$, and the sign function

in (3.3) can be predicted correctly with probability approaching one.

Useful byproducts of the partially linear additive quantile regression model are prediction intervals. A $(1-\alpha)\times 100\%$ prediction interval for next-year expenditure for a new patient with predictors X^* is $\left(\widehat{Q}_{Y^*|X^*}(\alpha/2),\ \widehat{Q}_{Y^*|X^*}(1-\alpha/2)\right)$. Though not necessary for classifying a future patient, the prediction interval provides useful information for the analyst.

Many statistical software packages such as `R`, `SAS` and `STATA` can be adapted for the first step of the prediction procedure. We recommend using the `R` package `plaqr` we developed (Maidman, 2017). A complete implementation of the prediction procedure using `plaqr` is given in Web Appendix D.

3.3. *Connection to the Binomial Regression Approach*

An alternative approach to this prediction problem relies on binomial regression with artificially dichotomized binary response variables. The underlying model with a logistic link function assumes that

$$\log\left(\frac{P(Y^*>c\mid X^*)}{1-P(Y^*>c\mid X^*)}\right)=X^{*'}\alpha$$

and with a complementary log–log link function that

$$\log\left\{-\log\left[1-P(Y^*>c\mid X^*)\right]\right\}=X^{*'}\alpha$$

for some unknown parameter α. In practice, α is usually estimated using the likelihood method to yield an estimator of the class probability $P(Y^*>c|X^*)$.

However, different from the ordinary binary classification problem for which only class labels are observed, in our setting, we also have complete information on the magnitude of the response variable. Our proposed semiparametric procedure fully uses the information in the response variable to make predictions. As binomial regression requires artificially dichotomizing the response variable, loss of information is expected.

4. Monte Carlo Studies

4.1. *Simulation Setup*

We compare our proposed new method (denoted by PLAQR) with five alternative parametric or semiparametric procedures, linear logistic regression (LLOG), partially linear additive logistic regression (PLALOG), linear complementary log–log regression (LCLOG), partially linear additive complementary log–log regression (PLACLOG), and the proposed prediction algorithm using classical linear quantile regression (LQR) (Koenker and Bassett, 1978), as well as a classification tree (TREE) in Monte Carlo experiments. The classification tree procedure incorporates different choices of r by treating the unequal cost of errors as a priori known class probabilities (Breiman et al., 1984) and is implementable in many software packages. For the binomial regression and classification tree approaches, the continuous response is dichotomized using a predetermined threshold c. The simulation results are based on 10,000 runs.

Mimicking the setting of the real data example in Section 5, we generate the response variable, next-year expenditure Y from the following model

$$Y=\exp\left(3V_1+1.5V_2+2V_3+b\left[\sin(2\pi Z_1)+Z_2^3\right]+\epsilon\right), \qquad (4.1)$$

where $V_1\sim$ Bernoulli(0.5), $V_2, V_3 \overset{iid}{\sim} N(0,1)$, $Z_1\sim$ Uniform (0,1), and $Z_2\sim$ Uniform$(-1,1)$. We consider three different choices for the random error distributions: (1) $\epsilon\sim N(0,1)$, (2) $\epsilon\sim t_3$, and (3) $\epsilon\sim V_2\,\xi$, where $\xi\sim N(0,1)$. Case (2) corresponds to a heavy-tailed error distribution, and case (3) corresponds to a heteroscedastic error distribution. We consider four different choices of b: 1, 2, 3, and 5, which provide varying magnitudes of nonlinearity. In each simulation scenario, the size of the training and testing data are both 200. A new patient is referred to as high-cost if his or her expenditure exceeds a threshold c. Here, we consider a choice of c corresponding to approximately the marginal 0.9 quantile of Y.

Step (1) of the prediction procedure described in Section 3.2 requires estimating a partially linear additive (or linear for LQR) quantile regression model from the training data. Geraci and Jones (2015) proposes a one-parameter symmetric monotonic transformation of the response to achieve linearity for $\mathbb{R}^+$ valued responses. In each iteration, we estimate the transformation parameter for each value value of τ. Letting $\tilde{Y}$ denote the transformed response, we estimate the quantile function for $\tilde{Y}$. To simplify computations, transformations for the PLAQR procedure are estimated for the model with three basis functions.

Motivated by Lee et al. (2014), we select the order of the basis functions m used to approximate each nonlinear component in the partially linear additive quantile regression model by minimizing

$$\mathrm{BIC}(m)=\log\left(\sum_{i=1}^{n}\rho_\tau\left(\tilde{Y}_i-\left[V_i'\hat{\beta}^{(m)}+\sum_{k=1}^{q}\pi(Z_{ik})^{(m)'}\xi_k^{(m)}\right]\right)\right)+(p+1+qm)\frac{\log n}{2n},$$

where the superscript (m) denotes estimates from the model with basis functions of order m.

4.2. *Simulation Results*

Different procedures are compared by plotting modified decision curves (see Vickers and Elkin (2006)). For each procedure and choice of r, let t_h and f_h denote the number of correctly and incorrectly predicted high-cost patients, respectively, and n denote the size of the prediction set. The decision curves are a plot of the net benefit for each prediction procedure:

$$\text{Net Benefit } = \frac{t_h}{n}-\frac{f_h}{n}r.$$

This measure reflects the simultaneous goals of achieving high sensitivity and high specificity by weighting the number of false positives by the relative cost of an error, r. Higher values indicate better prediction performance. We consider nine choices of r: $1, 9/11, 8/12, 7/13, 6/14, 5/15, 4/16, 3/17,$ and $2/18$ (corresponding to $\tau=0.50, 0.55, 0.60, \ldots, .90$, respectively), reflecting situations when the cost of misclassifying a high-cost patient is equal to nine times higher than misclassifying a not-high-cost cost patient.

We report the decision curves for $b = 5$ and the three types of errors in Figure 1. Due to space limitation, we report the decision curves for all other combinations of b and the three types of errors in Web Figure S1 in Web Appendix B. The decision curves all follow similar patterns as the curves in Figure 1. We summarize the major observations below.

First, we observe the importance of incorporating the nonlinear covariate effects. The more flexible semiparametric approach to classification outperforms the linear model based approaches and the classification tree, resulting in larger net benefit. Even when the nonlinear effects are milder (b=1 or 2), we observe the semiparametric models outperforming the linear models and classification tree. As the magnitude of nonlinearity increases, the increase in net benefit using the semiparametric approach becomes more evident.

Second, we observe that when the main interest is to predict if a future observation belongs to a small class, it is important to consider different weights for a false positive and a false negative in order to increase sensitivity. The increased sensitivity does not necessarily come at the cost of dramatically reduced specificity. When $r = 2/18$, our proposed new semiparametric procedure achieves a fine balance between sensitivity and specificity, resulting in the largest net benefit.

Finally, the most interesting and important observation is that PLAQR, PLALOG, and PLACLOG all perform similarly with respect to specificity; but PLAQR has higher sensitivity, particularly when $r = 2/18$. The poor performance of TREE can be explained by its low sensitivity for all choices of r, making it unusable in application. To better understand the relative performance of PLAQR versus PLALOG and PLACLOG, we consider a hypothetical situation in which 10,000 patients need to be classified as high-cost or not-high-cost, of which 1,000 are high-cost. When $b = 5$ with heteroscedastic errors and $r = 2/18$, PLAQR has mean sensitivity (SN) 0.981 and mean specificity (SP) 0.958, PLALOG has SN=0.864 and SP=0.964, and PLACLOG has SN=0.851 and SP=0.966. Translating these results into the above hypothetical setting, PLAQR predicts 19 false negatives and 378 false positives; while PLALOG predicts 136 false negatives and 324 false positives and PLACLOG predicts 149 false negatives and 306 false positives. Hence, applying PLALOG or PLACLOG results in 117 or 130 more high-cost patients falsely predicted as not-high-cost. With normal errors, applying PLALOG or PLACLOG results in 116 or 121 more high-cost patients being misclassified. With t_3 errors, applying PLALOG or PLACLOG results in 71 or 74 more high-cost patients being misclassified.

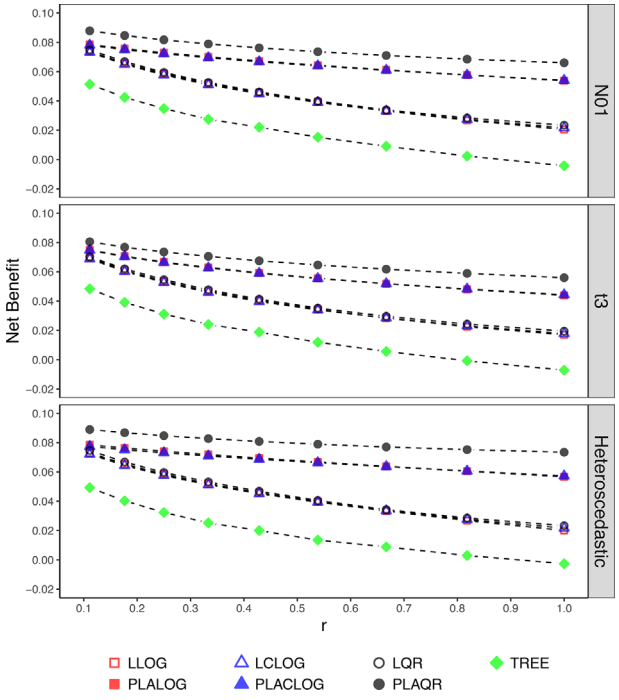

Figure 1. Decision curves for the LLOG, PLALOG, LCLOG, PLACLOG, LQR, TREE, and PLAQR procedures for simulations with standard normal errors, t_3 errors, and heteroscedastic errors when $b = 5$. All standard errors are less than 2.7×10^{-4}.

4.3. *Sensitivity Analysis*

In the following, we perform a sensitivity analysis to investigate the performance of the proposed semiparametric classification method when the underlying model does not have additive nonlinear effects. In particular, we consider responses generated from the log-linear model

$$Y = \exp\left(3V_1 + 1.5V_2 + 2V_3 + Z_1 + Z_2 + \epsilon\right),$$

and the model with nonadditive effects on the log scale

$$Y = \exp\left(3V_1 + 1.5V_2 + 2V_3 + Z_1 + Z_2 + Z_1Z_2 + \epsilon\right),$$

where the covariates, errors, training and testing data sample sizes, and number of iterations are the same as in Section 4.1. The log-linear model is a special case of the partially linear additive assumption while the nonadditive effects model violates it. We compare our proposed method with the correctly specified quantile based procedure (denoted ORACLE_QR). The one-parameter symmetric monotonic transformation is used to estimate transformations (Geraci and Jones, 2015). Decision curves are plotted in Figure 2.

When all effects are linear on the log scale, PLAQR and ORACLE_QR have almost identical estimated net benefits for all three errors. It is not surprising that PLAQR performs nearly as well as ORACLE_QR in this setting because the class of partially linear additive models contains the class of linear models. Even with nonadditive effects on the log scale, PLAQR only has slightly lower net benefit than ORACLE_QR. The results from this sensitivity analysis suggest that our proposed semiparametric classification procedure works well even when the model does not contain nonlinear effects or has nonadditive effects.

5. Analysis of Medical Expenditure Panel Survey

We now apply the proposed procedure to analyze medical expenditure from MEPS. Each panel consists of data from an individual over a two year span. In our analysis, we consider 1985 male patients aged 65 or older in Panels 1, 2, and 3 from years 2006-2007 (724 patients), 2007–2008 (568 patients), and 2008–2009 (693 patients), respectively. We use the data from

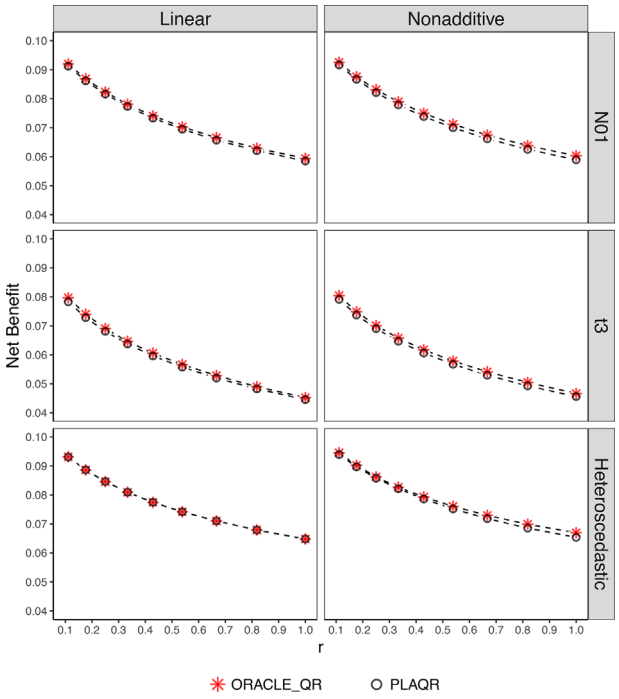

Figure 2. Decision curves for the ORACLE_QR and PLAQR procedures for log-linear and log-nonadditive model simulations with standard normal errors, t_3 errors, and heteroscedastic errors. All standard errors are less than 2.3×10^{-4}.

Panels 1 and 2 to predict if patients in Panel 3 will be high-cost in 2009. A threshold of US $28,520 corresponding to the marginal approximate 0.9 quantile of the next-year expenditure in Panel 3 is used to define patients as high-cost or not-high-cost.

Next-year expenditure among the three panels ranges from 0 to US $314,400 (mean and median are US $10,110 and US $3900, respectively). Nearly all of next-year expenditures are less than US $150,000 and about 4% of next-year expenditures are 0. The first and third quantiles are US $1539 and US $10,586, respectively. A histogram of next-year expenditure excluding the one expenditure greater than US $150,000 (to obtain sufficient resolution on the x-axis) is given in Figure 3. Its distribution is highly skewed. We use the following eight predictors observed in the first year of each panel: rgn (region of the country: northeast, midwest, south, west), insr (type of medical insurance: Medicare, private, Medicaid, uninsured), chrnc (number of chronic conditions: 0,1,...,8,9^+), prscrpt (number of prescriptions: 0,1,2,3,4^+), er (number of visits to the emergency room), health (summary score of self-described physical health), age, and rrs (relative risk adjustment score to account for inflation). The relative risk adjustment score, rrs, is a prospective measure of disease burden relying on health condition categories. Studies have shown that individuals with higher relative risk scores go on to use more hospital resources. These variables are important in the medical cost literature for their predictive power (Fleishman and Cohen, 2010).

First, we compare the prediction performance of the seven procedures LLOG, PLALOG, LCLOG, PLACLOG, TREE,

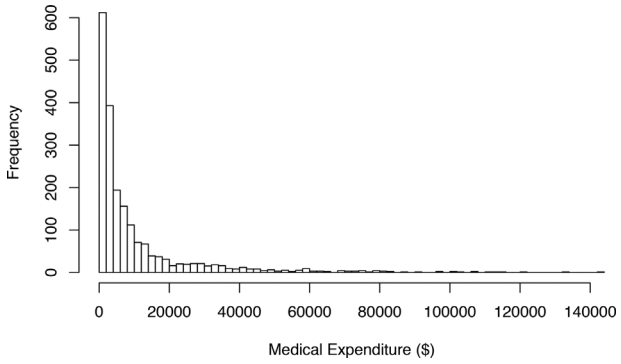

Figure 3. A histogram of next-year medical expenditures (second year of Panels 1, 2, and 3).

LQR, and PLAQR discussed in Section 4. For each of the seven procedures, we use the training data to fit the prediction model. To reflect the panel-to-panel changes in the next-year expenditure distribution and the goal of predicting patients with next-year expenditure greater than US $28,520 in Panel 3, we artificially dichotomize the next-year expenditure in Panels 1 and 2 according to their respective marginal approximate 0.9 quantiles (US $29,630 and US $24,000) for the binomial regression and TREE procedures. We assume nonlinear effects for age and rrs based on exploratory data analysis.

Transformations for the quantile regression procedures require strictly positive responses. Because some patients in the training data have 0 expenditure, we add 1 to each response and apply the recommended one-parameter symmetric transformation (Geraci and Jones, 2015) for each value of τ under consideration. The 95% bootstrap confidence intervals for the transformation parameters suggest that the transformation $\tilde{y}_i \equiv \log(y_i + 1)$ is appropriate for all quantiles. By the equivariance property of quantile regression, the conditional quantile of y is given by $Q_{Y|X,Z}(\tau) = \exp\left(Q_{\tilde{Y}|X,Z}(\tau)\right) - 1$.

We assess the overall lack-of-fit for the PLAQR model via the simulation based graphical method proposed by Wei et al. (2006). More specifically, we generate a random $\tilde{\tau}$ from the Uniform(0,1) distribution and estimate $\widehat{Q}_{Y|X}(\tilde{\tau})$ for a randomly sampled X in the training data. We repeat this process 5000 times to produce 5000 simulated responses from the assumed model and plot the quantiles of the sample responses against the quantiles of the simulated respones in Figure 4. The points in the QQ plot fall nearly along the identity line suggesting no lack-of-fit.

We evaluate the performance of all seven procedures for choices of r ranging from 1/9 to 1. When $r = 1$ ($\tau = 0.5$) none of the seven procedures is able to accurately predict high-cost patients. For smaller choices of r, the prediction procedures achieve a better balance of sensitivity and specificity. When $r = 1/9$ ($\tau = 0.9$), the procedures identify high-cost patients at an acceptable rate without sacrificing much ability to identify not-high-cost patients. Sensitivity from the PLACLOG procedure was 0.671 and from the PLAQR procedure 0.729. PLAQR is able to correctly predict 5.8% more high-cost patients than PLACLOG while maintaining adequate specificity. The specificities of PLACLOG and PLAQR were 0.713

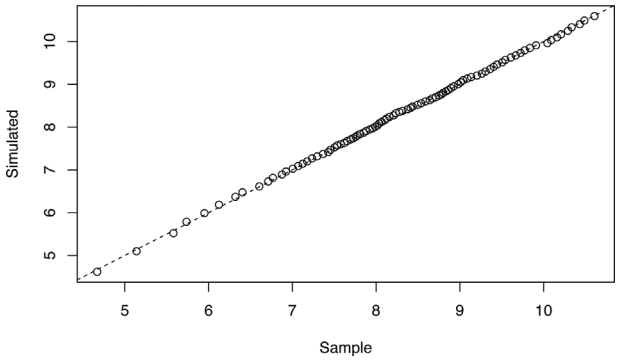

Figure 4. Lack-of-fit diagnostic QQ plot for PLAQR.

and 0.724, respectively. PLALOG had a sensitivity and specificity of 0.657 and 0.713, respectively. Consistent with findings in Section 4.2, the TREE procedure's low sensitivity rendered it inviable as a prediction procedure.

To better understand the practical importance of this increased sensitivity, consider that the subpopulation of males aged 65 and older in the U.S. in 2014 was about 20 million (U.S. Census Bureau, 2016). If about 10% of patients had high-cost medical expenditure, then PLAQR correctly identifies about 116,000 more high-cost patients than PLACLOG while correctly identifying slightly more not-high-cost patients.

Next, to gain more insight into this data, we further explore the estimated conditional 0.9 quantile of next-year expenditure in Panel 3 using the partially linear additive quantile regression model. The estimated coefficients for the linear effects and the estimated nonlinear functions $\hat{g}_1$ and $\hat{g}_2$ are given in Table 1 and Figure 5, respectively. Dashed lines are one standard deviation above and below the estimated effects. The pointwise standard deviations and confidence intervals are estimated from 999 bootstrapped samples using the wild bootstrap (Feng et al., 2011). Dashed lines in the plot of $\hat{g}_2$ do not cover the whole range of observed relative risk adjustment score due to sparsity in the large values of the observed relative risk adjustment score causing error estimation to be difficult and untrustworthy.

We conclude our analysis of MEPS by investigating prediction intervals. We computed and plotted 90% prediction intervals for patients' next-year expenditure in Panel 3 in Web Figure S2 in Web Appendix C. About 87% of the prediction intervals cover the true next-year expenditure for each patient. As an example, consider a typical patient with rgn = northeast, insr = Medicare, chrnc = 4, prscrpt = 3, er = 0, age = 75, rrs = 2.5, and health = 55. The 90% prediction interval of this patient's next-year medical expenditure is (US $2550, US $47,728) with a predicted 0.9 quantile of US $38,155.

6. Discussion

Motivated by a real data application to identify potential future high-cost patients, we propose a new semiparametric procedure to predict whether a new response falls in the tail of the response variable distribution. We prove that the proposed semiparametric procedure is a consistent estimator of the Bayes rule for classification while avoiding estimating the class probability. Empirically, we show that the proposed procedure outperforms popular binomial regression and classification tree based classification procedures. Furthermore, the semiparametric approach incorporates nonlinear covariate

Table 1
Coefficient estimates of linear effects for MEPS model when $\tau = 0.9$ *(90% confidence intervals in parentheses)*

Coefficient	Estimate	
(Intercept)	8.088	(7.128, 8.913)
rgn_{MW}	0.163	(−0.445, 0.124)
rgn_{S}	−0.048	(−0.337, 0.188)
rgn_{W}	−0.332	(−0.614, − 0.093)
$\text{insr}_{\text{prvt}}$	0.298	(−0.534, 0.727)
$\text{insr}_{\text{Mdcd}}$	−3.634	(−13.806, − 2.519)
$\text{insr}_{\text{unin}}$	0.683	(−0.904, 2.056)
chrnc_1	2.198	(1.532, 3.348)
chrnc_2	2.700	(1.681, 3.649)
chrnc_3	2.279	(1.606, 3.296)
chrnc_4	2.582	(1.834, 3.580)
chrnc_5	2.768	(2.060, 3.831)
chrnc_6	3.093	(2.359, 4.160)
chrnc_7	2.856	(2.030, 3.950)
chrnc_8	2.696	(1.993, 3.865)
chrnc_{9^+}	2.893	(2.101, 3.994)
prscrpt_1	−0.844	(−1.664, − 0.388)
prscrpt_2	−0.972	(−1.608, − 0.481)
prscrpt_3	−1.076	(−1.747, − 0.590)
prscrpt_{4^+}	−0.884	(−1.567, − 0.387)
er	0.079	(−0.045, 0.201)
health	−0.013	(−0.022, − 0.006)

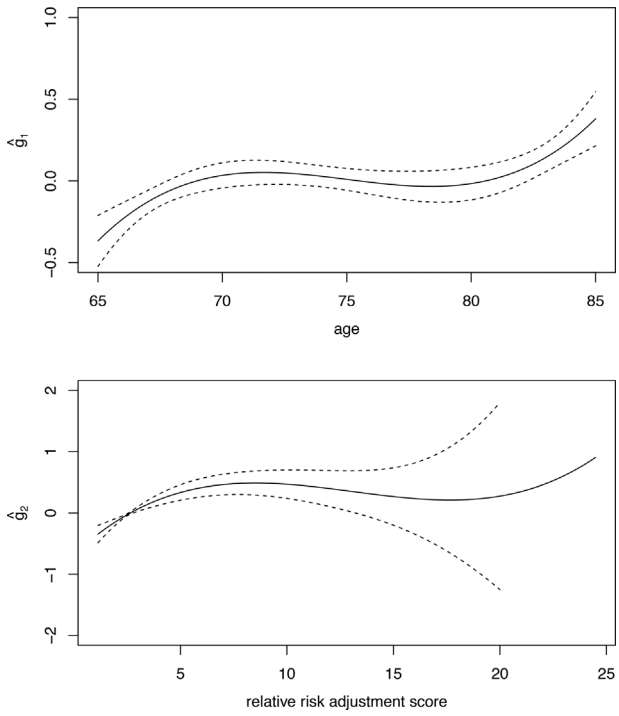

Figure 5. Plots of the estimated nonlinear effects.

effects. As suggested by simulation results, ignoring nonlinear effects may substantially increase the misclassification error rates.

In the real data application, we formulate the problem as a binary prediction problem as the intervention policy (whether to introduce an intervention program) only depends on whether the patient's future expenditure falls in the upper tail of the expenditure distribution. We then consider a decision theory framework to minimize the loss due to misclassification, where the two types of misclassification errors are weighted according to their potential consequences. If we can estimate the effect of the intervention as a percentage of the potential spending, then it is possible to formulate the decision theory framework as in Section 2.3 of Ehm et al. (2016) to take into account the magnitude of gains and losses. This approach will be useful in the future when information about medical expenditure reductions as a result of policy changes is available, for example, from a pilot program. This will be an interesting future research direction.

7. Supplementary Materials

Web Appendices and Figures referenced in Sections 3.2, 4.2, and 5, an R script implementing the proposed method, and the MEPS data analyzed in Section 5 are available with this article at the *Biometrics* website on Wiley Online Library. The `R` package `plaqr` is publicly available at `https://cran.r-project.org/package=plaqr`.

Acknowledgements

We thank the co-editor, the associate editor, and the anonymous referees for their helpful comments which helped us improve the article significantly. The research was partially supported by DMS-1712706 and a grant from the U.S. Department of Veterans Affairs.

References

Ahn, S., Smith, M. L., Altpeter, M., Post, L., and Ory, M. G. (2015). Healthcare cost savings estimator tool for chronic disease self-management program: A new tool for program administrators and decision makers. *Frontiers in Public Health* **3**, 42.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., et al. (2008). Algorithmic prediction of health-care costs. *Operations Research* **56**, 1382–1392.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Monterey, California: CRC press.

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **78**, 505–562.

Feng, X., He, X., and Hu, J. (2011). Wild bootstrap for quantile regression. *Biometrika* **98**, 995.

Fleishman, J. A. and Cohen, J. W. (2010). Using information on clinical conditions to predict high-cost patients. *Health Services Research* **45**, 532–552.

Geraci, M. and Jones, M. (2015). Improved transformation-based quantile regression. *Canadian Journal of Statistics* **43**, 118–132.

He, X. and Shi, P. (1996). Bivariate tensor-product b-splines in a partly linear model. *Journal of Multivariate Analysis* **58**, 162–181.

He, X., Zhu, Z.-Y., and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89**, 579–590.

Hosmer Jr, D. W. and Lemeshow, S. (2004). *Applied Logistic Regression*. New York: John Wiley & Sons.

Jones, A. M., Lomas, J., and Rice, N. (2015). Healthcare cost regressions: Going beyond the mean to estimate the full distribution. *Health Economics* **24**, 1192–1212.

Kannan, R., Vempala, S., and Vetta, A. (2004). On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)* **51**, 497–515.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge university press.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society* **46**, 33–50.

Koenker, R. and d'Orey, V. (1987). Algorithm as 229: Computing regression quantiles. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **36**, 383–393.

Lee, E. R., Noh, H., and Park, B. U. (2014). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association* **109**, 216–229.

Maidman, A. (2017). *plaqr: Partially Linear Additive Quantile Regression*. R package version 2.0 (available from `http://CRAN.R-project.org/package=plaqr`).

Meenan, R. T., O'Keeffe-Rosetti, M. C., Hornbrook, M. C., Bachman, D. J., Goodman, M. J., Fishman, P. A., et al. (1999). The sensitivity and specificity of forecasting high-cost users of medical care. *Medical Care* **37**, 815–823.

Schoenman, J. and Chockley, N. (2011). *Understanding us Healthcare Spending. Report*. Washington, DC: National Institute for Health Care Management Foundation.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. New York, NY, USA, John Wiley & Sons.

Sherwood, B. and Wang, L. (2016). Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics* **44**, 288–317.

U.S. Census Bureau (2016). *U.S. and World Population Clock*. `https://www.census.gov/popclock/`.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making* **26**, 565–574.

Wang, H. J., Zhu, Z., and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics* **37**, 3841–3866.

Wei, Y., He, X. (2006). Conditional growth charts. *The Annals of Statistics* **34**, 2069–2097.

Zhou, X.-H., Stroupe, K. T., and Tierney, W. M. (2001). Regression analysis of health care charges with heteroscedasticity. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **50**, 303–312.

Received January 2017. *Revised August* 2017.
Accepted October 2017.